Coreference in Prague Czech-English Dependency Treebank

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, Jiří Mírovský
Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800 Prague 1
{nedoluzko,mnovak,cinkova,mikulova,mirovsky}@ufal.mff.cuni.cz

Abstract

We present coreference annotation on parallel Czech-English texts of the Prague Czech-English Dependency Treebank (PCEDT). The paper describes innovations made to PCEDT 2.0 concerning coreference, as well as the coreference information already present there. We characterize the coreference annotation scheme, give the statistics and compare our annotation with the coreference annotation in Ontonotes and Prague Dependency Treebank for Czech. We also present the experiments made using this corpus to improve the alignment of coreferential expressions, which helps us to collect better statistics of correspondences between types of coreferential relations in Czech and English. The corpus released as PCEDT 2.0 Coref is publicly available.

Keywords: parallel corpus, bilingual coreference, alignment, Czech, English

1. Introduction

Over the last years, cross-lingual studies have been attracting a great deal of attention. Cross-lingual studies on discourse phenomena are no exception, mainly motivated by the task of machine translation, which still often overlooks phenomena beyond the sentence span. Corpora annotated with coreference are arguably a valuable resource for such studies. However, they are mostly monolingual. To our best knowledge, coreference-annotated parallel corpora are very rare, with ParCor 1.0 (Guillou et al., 2014), and the original release of Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012, PCEDT) being the main representatives.

In this paper, we present *Prague Czech-English Dependency Treebank 2.0 Coref* (Nedoluzhko et al., 2016, PCEDT 2.0 Coref), which comprises large-scale manual annotation of coreference links in a parallel corpus of Czech and English texts. Furthermore, it includes improvements to alignment of coreferential expressions. Both the coreference and the alignment annotation aim at serving as high-quality data for coreference-related studies on these languages. The annotation has been built upon the original release of PCEDT 2.0, a Czech-English parallel treebank sized over 1.2 million tokens in almost 50,000 sentence pairs. Its English part consists of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1999), while the Czech part was manually translated from the English source sentence by sentence.

PCEDT is the second Praguian corpus with manual coreference annotation. It was preceded by the monolingual Prague Dependency Treebank (PDT), which pioneered the coreference annotation on Czech texts in its version 2.0 (Hajič et al., 2006), later enriched by other types of coreference and discourse relations in version 3.0 (Bejček et al., 2013). The differences between the PCEDT 2.0 Coref and PDT 3.0 coreference annotation are addressed in Section 3.5.

The annotation of coreference relations in PCEDT has proceeded in multiple stages. While the version 2.0 introduced by Hajič et al. (2012) contained the annotation of the so-called grammatical coreference and pronominal coreference, it has been recently enhanced by adding nominal coreference. Both annotation stages are elaborated in Sections 3.1. and 3.3.

The coreference annotation on the English part of PCEDT was built above an automatic transformation of the original coreference annotation extracted from the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005, BBN-PCETC) described in the Ontonotes coreference guidelines (BBN Technologies, 2006). It was further manually checked and corrected. Some distinctions between the PCEDT and Ontonotes coreference are described in Section 3.4.

This work also concentrates on the issue of cross-lingual alignment of coreferential expressions. As the standard unsupervised techniques for word-alignment are known to fall short of performance on function words and pronouns, we introduce a supervised approach exploiting the manually annotated alignment by Novák and Nedoluzhko (2015). The details of the method are spelled out in Section 4.

PCEDT 2.0 Coref together with its documentation is publicly available[1]. The same holds for the source code used to prepare the corpus and collect the statistics presented in this paper.[2]

2. Related corpora

Parallel corpora with coreference annotation are very rare. The Romanian-English corpus introduced by Postolache et al. (2006) is probably the first one with the coreference annotation. The corpus, sized over 600 sentences, was manually annotated with full-fledged coreference chains in both languages. However, it is not publicly available.

To our best knowledge, besides PCEDT, only ParCor 1.0 (Guillou et al., 2014) belongs to the category of publicly available coreference-annotated parallel corpora. It is a German-English parallel corpus presenting manual annotation of more than 8,000 sentences. Unlike PCEDT, texts

[1] http://ufal.cz/pcedt2.0-coref

[2] https://github.com/ufal/pcedt2.0-coref

in the corpus come from two different genres: transcribed planned speech from TED Talks, and written texts from EU Bookshop. On the other hand, only pronominal coreference is annotated; it does not take into account full referring expressions.

Some of the corpora contain automatic coreference annotation. CzEng 1.0 (Bojar et al., 2012) is a large-scale Czech-English parallel corpus consisting of more than 15 million sentence pairs from several different domains. The automatic coreference annotation has been obtained independently for each of the languages. It includes pronominal and zero coreference according to the Prague coreference annotation tradition (see Section 3.1. below).

3. Bilingual coreference annotation in PCEDT

The Prague Czech-English Dependency Treebank is a parallel corpus annotated at several layers of linguistic representation up to the layer of deep syntax (or *tectogrammatical layer*), drawing on the Functional Generative Description (Sgall, 1967; Sgall et al., 1986). The tectogrammatical representation of a sentence is a dependency tree with semantic labeling, coreference and argument structure description based on a valency lexicon. The nodes of a dependency tree are formed only by auto-semantic words (with some exceptions of a technical nature). Furthermore, some expressions that are absent on the surface are reconstructed at the tectogrammatical layer. For example, in PCEDT, anaphoric zeros are introduced in the tectogrammatical layer with a newly established node, e.g. elided subjects in Czech in Example (1).

(1) *Nepřišel. [0(elided he) Didn't come.]*

3.1. Coreference relations: annotation scheme

The coreference annotation of PCEDT takes place on the tectogrammatical layer to allow the marking of zero anaphora. The annotation covers the cases of grammatical (syntactic) and textual coreference. In the following, each coreference relation consists of two arguments: the *anaphor* is a referring expression, and the *antecedent* is the expression which it refers to.

Grammatical coreference. The grammatical coreference typically occurs within a single sentence, the antecedent is expected to be derived on the basis of grammar rules of a given language. It concerns the following cases:

- relative pronouns (*Alex is* ***the boy who*** *kissed Mary*);
- the arguments of the verbs of control (***Peter*** *wants [***Ø*** *to sleep*].);
- reflexive pronouns (***My daughter*** *likes to dress* ***herself*** *without my help*);
- coreference of arguments 'hidden' in reciprocal constructions (***Peter*** *_i and* ***Mary*** *_j kissed* ***Ø*** *_i+j* .);
- coreference with verbal modifications that have dual dependency (*John saw* ***Mary*** *[***Ø*** *run around the lake]*).

Textual coreference. In this type of coreference, arguments are not realized by grammatical means alone, but also via context. Anaphoric (occasionally cataphoric) referential devices are expressed by various language means (pronouns, synonyms, generalizing nouns, etc.).[3] Within textual coreference, we annotate the following types:

- Pronominal coreference with personal, possessive and demonstrative pronouns (Example (2));

(2) ***A form of asbestos once used to make Kent cigarette filters*** *_i has caused a high percentage of cancer deaths among a group of workers exposed to* ***it*** *more than 30 years ago, researchers reported. [Výzkumníci uvedli, že* ***forma azbestu kdysi používaná k výrobě cigaretových filtrů značky Kent*** *způsobila vysoký podíl úmrtí na rakovinu mezi dělníky, kteří* ***jí*** *byli vystaveni před více než 30 lety.]*

- Coreference with textual ellipsis. In this case, a new node with the lemma substitute `#PersPron` is added to the tectogrammatical tree. Textual ellipsis is especially frequent in Czech (see Section 3.), but it is also common in some syntactic constructions in English. For instance, in Example (3) with a coordinative construction of an active and a passive clauses, the unexpressed patient argument of the verb *reject* is reconstructed at the tectogrammatical layer, according to the valency lexicon.[4] The coreference link is annotated to the subject of the active part of a coordinative pair.

(3) ***More common chrysotile fibers*** *are curly and are more easily* ***Ø*** *rejected by the body, Dr. Mossman explained.*

- Nominal textual coreference. We do not annotate anaphoric relations in a restricted sense, but we concentrate on marking the equivalence of referents of antecedent and anaphoric expressions. For instance, in Example (4), coreference is marked for the relation between *Fujitsu Ltd.* and *Japan's biggest computer maker*, although in the English original text, the noun phrase *Japan's biggest computer maker* contains no explicit anaphoric reference to the antecedent *Fujitsu Ltd.* It is interesting, however, that in the Czech translation, the anaphoric reference is used (*Tento největší počítačový výrobce v Japonsku* [lit. *This Japan's biggest computer maker*]).

(4) *Japanese companies have long been accused of sacrificing profit to boost sales. But* ***Fujitsu Ltd.*** *has taken that practice to a new extreme.* ***Japan's biggest computer maker*** *last week*

[3] The detailed description of the distinction between the grammatical and textual coreference can be found e.g. in (Mikulová et al., 2006).

[4] The valency lexicons for Czech (PDT-Vallex) and English (Engvallex) have been comprehensively described in (Hajič et al., 2003) and (Urešová et al., 2015), respectively.

*undercut seven competitors to win a contract to design a mapping system for the city of Hiroshima's waterworks. [Japonské společnosti jsou již dlouho obviňovány z toho, že se vzdávají zisku, aby zvýšily obrat. Ale **firma Fujitsu Ltd.** tuto praxi dovedla do nového extrému. **Tento největší počítačový výrobce v Japonsku** minulý týden nabídl nejnižší cenu v porovnání se svými sedmi konkurenty a získal kontrakt na projekt mapovacího systému pro zásobování města Hirošimy vodou.]*

The textual coreference is marked up to the length of 20 sentences. Annotating coreference for a greater number of sentences is possible only in cases of automatic pre-annotation of named entities coreference. This decision was made in order to avoid a large number of mistakes and to reach higher inter-annotator agreement.

Special cases of textual coreference. In accordance with the Prague coreference annotation tradition, two special cases of reference are annotated in PCEDT. First, we mark the cases of endophoric *references to a discourse segment* of more than one sentence, including the cases where the antecedent is understood by inference from a broader co-text. This kind of relation has no explicitly marked antecedent; it just proves the fact that the given anaphoric nominal group co-refers with some discourse antecedent of more than one sentence. Second, a specifically marked link for *exophora* denotes that the referent is "out" of the co-text;i.e., it is only known from the actual situation.[5] Exophoric reference is annotated in case of temporal and local deixis (*this year, this country*), deixis with pronominal adverbs (*here*), as well as exophoric reference to the whole text (e.g. *this report* referring to the whole actual report in Example (5)), etc.

(5) *The information in **this report** is a free service to businessmen. [Informace v **tomto přehledu** jsou bezplatnou službou podnikatelům.]*

To develop a maximally consistent annotation scheme, we follow a number of basic principles, such as the principle of the maximum length of coreferential chains, the principle of maximal size of an anaphoric expression (subject to annotation is always the whole subtree of the antecedent/anaphor), and the principle of cooperation with the syntactic structure of a given dependency tree, which does not let us annotate relations which are already caught up by the syntactic structure of the tectogrammatical tree.

3.2. Coreference chains

The annotation of textual coreference in PCEDT is based on the chain principle, the anaphoric expression always referring to the last preceding coreferential antecedent. There may be only one textual coreference arrow leading from/to a tectogrammatical node. No ambiguity is annotated. If noted by annotators, the most likely variant had to be chosen, other options were marked as comments to coreference arrows. The coreferential chain formed by this principle represents an *entity* and the individual coreferring items are denoted as *mentions*. The very first mention of each entity is thus the only mention in the coreference chain that contains no outgoing coreference link, unless it is an exophora or reference to a segment.

In case of textual reference to multiple antecedents (so-called *split antecedent*), coreference relations of a special type were annotated.[6] This is the case when the anaphor C is coreferential with the union of antecedents $A \cup B$, both present in tectogrammatical structure of the corresponding text. For example, the anaphoric nominal group *the companies* in the third sentence in Example (6) refers to two antecedents, i.e. to *Cray Research* and *Cray Computer*.

(6) *Under terms of the spinoff, **Cray Research** stockholders are to receive one **Cray Computer** share for every two Cray Research shares they own in a distribution expected to occur in about two weeks. No price for the new shares has been set. Instead, **the companies** will leave it up to the marketplace to decide.*

For grammatical coreference, multiple arrows are allowed when an anaphoric expression (mostly a reconstructed anaphoric zero or a relative pronoun) refers to more than one antecedent.

3.3. Annotation process

The coreference annotation in PCEDT has been completed separately and independently for the Czech and the English sides. Furthermore, it proceeded in two stages.

During the first stage, grammatical coreference for both Czech and English was automatically pre-annotated by heuristics and manually corrected according to annotation guidelines in (Mikulová et al., 2006). The automatic annotation had been trained on the Czech data from the Prague Dependency Treebank (Bejček et al., 2013), with grammatical coreference previously annotated and corrected. Next, pronominal coreference in Czech was manually annotated. As for the English pronominal coreference, its annotation was built upon an automatic transformation of the original coreference annotation extracted from the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005, BBN-PCETC). It was further manually checked and corrected. Together with the correction, coreferential links coming from the elided nodes were added.

The second stage comprised the annotation of nominal coreference. For English, a part of nominal coreferential links has been extracted from what was in that time accomplished in BBN-PCETC. It accounts for ca 20% of nominal coreferential links. The rest had to be annotated manually, using the Ontonotes coreference guidelines (BBN Technologies, 2006). For the Czech part of the PCEDT, nominal textual coreference has been annotated manually.

[5] We are aware of the fact that the term *coreference* is usually used only for endophoric reference but still the annotation of exophoras is technically included into the coreference annotation.

[6] Technically, this is the same type of arrows as we used for the annotation of bridging set–subset relations in PDT (Nedoluzhko et al., 2013).

All the annotations were produced by a group of human annotators, students of linguistics. Each text was annotated by one annotator; small parts of the data were annotated in parallel by two annotators to measure the inter-annotator agreement. The F1-measure of inter-annotator agreement varies within 75-85% for pronominal coreference, and within 70-80% for nominal coreference, depending on the length and abstractness degree of the texts.

From a technical point of view, we used a special extension of a highly customizable tree editor – TrEd (Pajas and Štěpánek, 2008) – for coreference annotation. Visualized by this editor, each of the annotated relations is represented as an arrow connecting two tectogrammatical nodes. Types of relations are specified in special attributes. Figure 1 shows an example sentence pair visualized in TrEd, with annotation of coreference highlighted in solid green.

3.4. Relation of coreference annotations in PCEDT and Ontonotes

The coreference annotation in the English part of PCEDT is closely related to the annotation of coreference in BBN-PCETC (BBN Technologies, 2006). As described in Section 3.3., the textual pronominal coreference annotation is based on BBN-PCETC, and the textual nominal coreference annotation continues the annotation which was applied on a piece of the data there. During the annotation of the rest of the data with nominal coreference, the Ontonotes coreferential links were once more manually checked and corrected. The changes are mainly technical. Ontonotes allow for multiple links from a single node which may have different reasons (split antecedents, ambiguity, referring both to the apposition/coordination constructions and to the members of the constructions) - in PCEDT these multiple links have been classified and changed according to the coreference annotation guidelines described in Section 3.1.. Some minor annotation mistakes have been corrected, too. Moreover, coreference links that appeared due to the reconstruction of the tectogrammatical structure of the sentences have been added.

Another difference is the absence of reference to larger textual segments in the Ontonotes coreference annotation (reference to the nearest possible antecedent has been used instead). For such cases, we added reference to larger textual segments in PCEDT.

The Ontonotes convention was not to annotate coreference relations with indefinite and non-specific nominal groups. In PCEDT, we did our best to keep this convention. However, in some cases coreferential relations between nominal groups without definite determiners are so obvious that they were annotated even in Ontonotes, in spite of negative instructions (as in Example (7)).

(7) *China's parliament ousted two Hong Kong residents from a panel drafting* ***a new constitution*** *for the colony. [...] The committee is formulating* ***Hong Kong's constitution*** *for when it reverts to Chinese control in 1997... [Čínský parlament vyloučil dva obyvatele Hong Kongu ze skupiny odborníků, která má vytvořit návrh* ***nové ústavy*** *této kolonie. [...] Výbor formuluje* ***ústavu****, kterou se Hong Kong bude řídit, až v roce 1997 přejde pod čínskou správu...]*

3.5. Relation of coreference annotations in PCEDT and PDT

The grammatical and pronominal textual coreference in PCEDT was annotated according to the same guidelines as in PDT (Mikulová et al., 2006). As for the nominal coreference, the guidelines had to be simplified to preserve the correspondence with the Ontonotes coreference (see Section 3.4.). For instance, only nominal groups with specific reference have been annotated for coreference in PCEDT, as opposed to PDT 3.0, where the annotation includes also generics, distinguished from coreference of specific nouns by a special attribute (Nedoluzhko et al., 2013). Nevertheless, we realize that avoiding nominal groups with generic reference is problematic, especially in languages without the grammatical category of definiteness. For example, in Czech, there are non-obligatory formal signals for specific definite nominal groups. Thus, in some cases it is difficult to decide whether a given nominal group should be annotated for coreference. On the other hand, generic nouns may be used anaphorically and with a determiner (as in Example (8)) inciting annotators to mark a coreferential relation, regardless the annotation instructions. Also, the distinction between generic and specific nominal groups is often ambiguous.[7]

(8) *The sterilizing gene is expressed just before the pollen is about to develop and it deactivates the anthers of every flower in* ***the plant****. Mr. Leemans said this genetic manipulation doesn't hurt the growth of* ***that plant****. [Sterilizační gen se projeví těsně předtím, než se pyl začíná vytvářet, a zneškodní prašníky každého květu* ***rostliny****. Leemans řekl, že tento genetický zásah neohrožuje růst* ***rostliny****.]*

One more significant difference between the annotations of PCEDT and PDT 3.0 is that bridging relations were not included in PCEDT, except for a special case of split antecedents described in Section 3.2..

3.6. Statistics

Table 1 presents the statistics of coreferential relations annotated in the PCEDT.

As observed from the table, the total number of coreferring nodes in both languages is comparable. Also the number of textual nominal coreference is similar. The main difference concerns the numbers for the grammatical and textual pronominal coreference in English and Czech. The table shows that while English has significantly more grammatical coreference links, Czech uses textual pronominal coreference more often. There is a set of reasons for this difference. Mostly, it concerns the preference of different types of constructions in English and Czech. For example, in English, infinitive, participle and gerund clauses are commonly used, especially in newspaper texts that are subject

[7] For a more detailed analysis of the challenge of coreference annotation with generics see e.g. (Friedrich et al., 2015) and (Nedoluzhko, 2013).

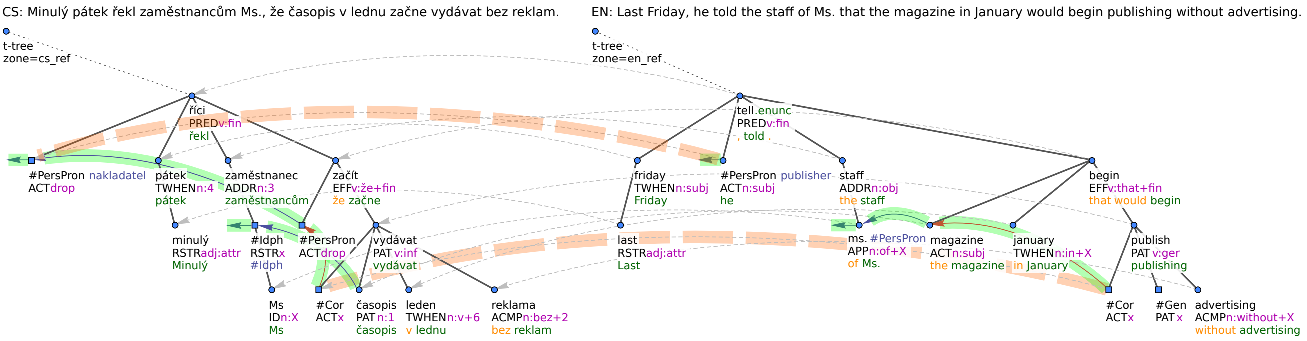

Figure 1: Example sentence pair from PCEDT. Coreference links are highlighted in solid green; alignment for coreferential expressions focused with the supervised approach is highlighted in dashed orange.

	English	Czech
Sentences	49,208	49,208
Tokens	1,173,766	1,151,150
Tectogrammatical nodes	838,212	931,846
Coreferring nodes	188,528	183,134
grammatical coreference	38,860	27,809
textual prononimal coreference	35,026	46,794
textual nominal coreference	56,377	51,839
first mentions	53,562	51,348
reference to split antecedents	455	577
reference to a segment	1,966	2,669
exophora	2,282	2,098
Entities	58,270	56,692
Non-singleton entities	54,190	52,027

Table 1: Statistics of the annotated data

of our analysis. These constructions are often translated as finite subordinate clauses into Czech. As the result, according to the definition of grammatical and textual coreference, in Example (9), we have grammatical coreference with the argument (reconstructed in the tectogrammatical structure) of the participle *accusing* in English. In Czech, the zero subject of the subordinate relative clause *ve které viní Darmana ze zaprodanosti* [lit. *in which Ø(he) accuses Mr. Darman of selling out*] co-refers with the subject of the main clause and this is the case of textual coreference.

(9) *He left a message accusing Mr. Darman of selling out. [Zanechal mu zprávu, ve které viní Darmana ze zaprodanosti.]*

One the other hand, it is worth noting that grammatical coreference rules for English and Czech are similar but not totally identical. The English grammar does not require that the argument of the participle in such cases occupying the semantic role of Actor be coreferential with the Actor of the governing node. For example, in the sentence above, both *he* and *message* could be the subject of the participle clause. Thus, strictly speaking, this case cannot be unambiguously considered to be grammatical coreference. Table 1 also shows a number of entities. Due to a chain principle, it is a sum of first mentions, reference to split antecedents,[8] exophoras, and references to a segment. However, some of these mentions may be *singletons*, i.e. they do not form a chain with any other mention. Therefore, we present the number of entities consisting of more than one mention (non-singleton entities), and include only these entities in the statistics of alignment in Section 4.

4. Improved alignment of coreferential expressions

An essential part of each parallel corpus is the alignment between languages. Except for sentence alignment, texts in PCEDT are also aligned on the word level. Originally, the word alignment in PCEDT was produced in an unsupervised way by the GIZA++ tool (Och and Ney, 2000). GIZZA++ was run in both directions; then, symmetrization of the two produced alignments was taken.[9] The alignment on the tectogrammatical layer, that is, the layer containing coreference annotation, was obtained by projecting the word alignment. Furthermore, a simple heuristics was applied for elided subjects reconstructed in the tectogrammatical tree.

The unsupervised approach of word alignment seems to be working sufficiently well for most of the words, especially for auto-semantic words. Nonetheless, its performance falls behind for words such as pronouns, not to speak of zeros unexpressed on the surface. These expressions play a key role in coreference relations. The problem is that such words are tightly associated with the grammar of a particular language, which usually differs across distant languages (see the discussion on different kinds of constructions in Section 3.6.).

Manually aligned data. Recently, Novák and Nedoluzhko (2015) presented a study of correspondences between mentions in Czech and English. For this purpose, they manually annotated selected coreferential expressions (central pronouns,[10] relative pronouns and

[8] If an anaphor $A \cup B$ refers to the antecedents A and B, we count it as 3 separate entities. If A does not co-refer with anything else, it is the only mention of its entity and we count it as a first mention.

[9] A union of two symmetrization approaches has been used: *intersection* and *grow-diag-final-and*.

[10] A term coined by Quirk et al. (1985) encompassing personal, possessive and reflexive pronouns.

	English				Czech			
	central	relative	zero	total	central	relative	zero	total
# occurrences	549	223	703	1,475	286	335	850	1,471
original	80.3	96.9	75.8	80.7	88.1	67.2	78.7	77.9
supervised	90.9	96.4	80.7	**86.9**	92.7	83.6	87.0	**85.6**

Table 2: Number of occurrences of coreferential expressions in the training data. The original and supervised method for aligning coreferential expressions measured by accuracy.

anaphoric zeros[11]) with their counterparts in the other language in 1,078 sentence pairs from PCEDT sections wsj_1900–49. The numbers of occurrences of particular mention types are shown in Table 2.

In sections wsj_1900–49 of PCEDT 2.0 Coref, this manual alignment annotation (labeled as coref_gold) replaces the original alignment, and the targeted coreferential nodes are indicated by the align-coref attribute. For all other nodes in these sections, the original alignment remains unchanged. In the example sentence pair in Figure 1, the manually annotated alignment links are highlighted in dashed orange.

Supervised resolver. We take the PCEDT sections with manual alignment as a training data and create two supervised resolvers for the selected coreferential expressions, one for English, the other one for Czech.

We used practically the same feature set as Novák and Žabokrtský (2014) employed for English central pronouns, i.e., the original GIZA++ alignment, graph features, grammatical features, and their combinations. Moreover, the features are combined with the type of the coreferential expression, i.e. a central pronoun, a relative pronoun, or an anaphoric zero, to capture potential differences between them.

The task is modeled as ranking all nodes from the aligned sentence to find a best-fitting candidate to a particular coreferential expression. A special candidate is included to comprise the option that the expression has no counterpart in the other language. The aligners have been trained using Stochastic Gradient Descent with L2 regularization in the Vowpal Wabbit[12] machine learning toolkit. In the resolution time, the aligners are applied to the targeted nodes, producing alignment links in both directions. They are subsequently symmetrized with a preference for links selected by both aligners. The remaining links are included in such a way that each targeted node is covered at most once.

The resulting supervised aligner has been run over all targeted nodes in PCEDT 2.0 Coref (again, indicated by the align-coref attribute), except for the manually annotated part. As with the manual alignment, the alignment annotation produced by the supervised approach (labeled as coref_supervised) replaces the original alignment. For all the other nodes, the original alignment remains unchanged.

Evaluation and analysis. To evaluate our supervised method, we carried out 10-fold cross-validation over the training data. The quality of the supervised alignment and the original PCEDT alignment in terms of accuracy is shown in Table 2. The supervised approach outperforms the method producing the original alignment for every type of coreferential expressions, except for the English relative pronouns, where the difference is marginal, though. Overall, the presented method exhibits the improvement of over 6% points.

Apart from the intrinsic evaluation, we also used an approximate approach allowing for large-scale evaluation, based on the following assumption: coreference is one of the means to maintain coherence in the text. If we assume that text coherence is not violated during translation, coreference chains representing an entity in each of the languages should correspond. Since language grammar differences have apparent effects on coreference,[13] this is far from being true. Nevertheless, alignment improvements should lead to a higher rate of entity correspondence.

We measure this tendency by two scores: the *coreferring counterpart ratio*, and the *entity alignment rate*. The former is calculated as a proportion of the coreferring nodes targeted by the supervised aligner, whose counterparts in the other language are also coreferring. On the other hand, the latter score takes all mentions into account. It takes the proportion of the nodes belonging to a single entity whose counterparts also belong to the same entity and averages it over all non-singleton entities. To gain a better insight into the alignment quality, we also observe the frequency of 1:0, 1:1 and 1:N entity mappings. The ideal situation would be if the 1:0 and 1:N mappings were rare, meaning that almost all entities correspond to exactly one counterpart entity in the other language.

Table 3 shows the scores measured on the wsj_1900–49 subset, as well as on the complete PCEDT 2.0 Coref, using three different alignment approaches for the nodes targeted by the supervised aligner: original, supervised, and manual.[14] As for the coreferring counterpart ratio and the entity alignment rate scores, substantially higher scores are observed for the manual alignment than for the original

[11] Czech zeros consist mainly of subjects dropped from the surface, while English zeros are usually unexpressed arguments of infinitives, *-ed* and *-ing* participles.

[12] https://github.com/JohnLangford/vowpal_wabbit

[13] Many examples for English and Czech can be found in (Novák and Nedoluzhko, 2015).

[14] As the annotation of manual alignment is missing outside the sections wsj_1900–49, this approach is not evaluated for the complete PCEDT.

	English						Czech					
	orig		super		manual		orig		super		manual	
Coref. counter. ratio	57.4	55.3	66.4	62.2	71.3	—	54.9	55.2	62.9	62.4	68.5	—
Entity alignment rate	56.2	49.7	59.3	52.0	60.5	—	53.4	52.2	56.1	54.9	57.8	—
1:0 entity mappings	24.8	30.9	23.5	29.8	22.8	—	28.7	28.1	26.9	26.8	26.0	—
1:N entity mappings	6.2	5.7	5.6	5.7	6.0	—	6.5	6.1	6.3	6.1	6.4	—

Table 3: The coreference-based metrics showing the quality of node alignment (in %), comparing the original, supervised and manual alignment. In each cell, the first number is measured on the sections wsj_1900–49, while the second one on the complete PCEDT 2.0 Coref.

alignment. This observation supports the aforementioned assumption. The numbers on supervised alignment accord with the scores from the intrinsic evaluation, performing better than the original alignment overall. A larger difference in both scores between original and supervised alignment for the wsj_1900–49 subset can be justified by subtle overfitting, as this dataset actually serves as the training data for the supervised method.

The proportion of entities with the 1:0 entity mapping drops with improving alignment, whereas the proportion of entities with the 1:N mapping exhibits no specific behavior. It suggests that the proposed alignment improvements stem mostly from adding new links than from modifying the existing ones. As the occurrence of entities with no counterparts may be, among other reasons, attributed to grammatical differences, such cases can hardly disappear. On the other hand, around 6% of the entities with the 1:N mapping must be a result of an error, either in alignment or coreference.

5. Conclusion

To summarize, as a result of this work we gained a fully manually annotated coreference corpus with a higher-quality word alignment of almost 50,000 English and Czech parallel aligned sentences. The developed data make it possible to attain a deeper understanding of anaphoric relations in English in comparison to Czech and vice-versa. Furthermore, we believe that analyzing a language from a multilingual perspective is not only beneficial with regard to cross-lingual tasks, but it also helps to understand various phenomena in that individual language in greater depth.

6. Acknowledgments

This project has been funded by the Czech Science Foundation grants GA-16-05394S and GA-15-20031S, the GAUK grant 338915, and the 7th Framework Programme of the EU grant QTLeap (No. 610516). This work has been also supported and has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project No. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.

7. Bibliographical References

BBN Technologies, (2006). *Co-reference Guidelines for English OntoNotes*.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., and Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey. European Language Resources Association. URL: http://hdl.handle.net/11234/1-1458.

Friedrich, A., Palmer, A., Sørensen, M. P., and Pinkal, M. (2015). Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop (LAW IX)*, Denver, Colorado, US. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association. URL: http://opus.lingfil.uu.se/ParCor.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, Vaxjo, Sweden. Vaxjo University Press. URL: http://hdl.handle.net/11858/00-097C-0000-0023-4338-F.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report TR-2006-30, Prague, Czech Republic.

Nedoluzhko, A., Mírovský, J., and Novák, M. (2013). A Coreferentially annotated Corpus and Anaphora Resolution for Czech. In *Computational Linguistics and Intellectual Technologies*, Moscow, Russia. ABBYY.

Nedoluzhko, A. (2013). Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, Sofija, Bulgaria. Omnipress, Inc.

Novák, M. and Nedoluzhko, A. (2015). Correspondences between Czech and English Coreferential Expressions. *Discours: Revue de linguistique, psycholinguistique et informatique*, 16.

Novák, M. and Žabokrtský, Z. (2014). Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pajas, P. and Štěpánek, J. (2008). Recent Advances in a Feature-rich Framework for Treebank Annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Postolache, O., Cristea, D., and Orasan, C. (2006). Transferring Coreference Chains through Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. European Language Resources Association.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, London, United Kingdom.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, Netherlands.

Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., and Šindlerová, J. (2015). Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus. In *Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015)*, Stroudsburg, PA, USA. Association for Computational Linguistics. URL: `http://hdl.handle.net/11234/1-1512`.

8. Language Resource References

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). *Prague Dependency Treebank 3.0*. Charles University in Prague, ÚFAL, URL: `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., and Urešová, Z. (2006). *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, URL: `http://hdl.handle.net/11858/00-097C-0000-0001-B098-5`.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). *Prague Czech-English Dependency Treebank 2.0*. Charles University in Prague, ÚFAL, URL: `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`.

Marcus, M., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1999). *Penn Treebank 3*. Linguistic Data Consortium, URL: `https://catalog.ldc.upenn.edu/LDC99T42`.

Nedoluzhko, A., Novák, M., Cinková, S., Mikulová, M., and Mírovský, J. (2016). *Prague Czech-English Dependency Treebank 2.0 Coref*. Charles University in Prague, ÚFAL, URL: `http://hdl.handle.net/11234/1-1664`.

Weischedel, R. and Brunstein, A. (2005). *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, ISLRN: 141-282-691-413-2.